University of Southern Denmark
IMADA
Arthur Zimek
Jonatan Møller Gøttcke, Jonas Herskind Sejr

DM566/DM868/DM870: Data Mining and Machine Learning

Spring term 2019

Exercise 4: Apriori, Color Histograms

Exercise 4-1 Apriori

Consider the following transaction database D over the items $I = \{A, B, C, D, E, F\}$.

TransID	Items
1	A B E
2	B D
3	C D F
4	A B D
5	A C E
6	B C E F
7	A C E
8	A B C E
9	A B C D F
10	B C D E

Given the support threshold $\sigma = 2$, apply the Apriori algorithm and extract all frequent itemsets w.r.t. the given threshold. Please explain in the solution all the steps that you followed.

In particular include for each level the candidate set (C_k) (i) after the join step before pruning and (ii) after pruning. Annotate for those objects pruned in (ii) the explicit reason for pruning them.

Also give explicitly the solution of frequent k-itemsets (S_k) for each k.

Exercise 4-2 Color-Histograms and Distancefunctions

As a warm-up on distance measures: For each of the following distance measures (Euclidean, Manhattan, maximum, weighted Euclidean, quadratic form)

$$\begin{aligned}
\mathrm{dist}_2(p,q) &= \left(|p_1-q_1|^2+|p_2-q_2|^2+|p_3-q_3|^2\right)^{\frac{1}{2}} \\
\mathrm{dist}_1(p,q) &= |p_1-q_1|+|p_2-q_2|+|p_3-q_3| \\
\mathrm{dist}_\infty(p,q) &= \max(|p_1-q_1|,|p_2-q_2|,|p_3-q_3|) \\
\mathrm{dist}_w(p,q) &= \left(w_1|p_1-q_1|^2+w_2|p_2-q_2|^2+w_3|p_3-q_3|^2\right)^{\frac{1}{2}} \\
\mathrm{dist}_M(p,q) &= \left((p-q)M(p-q)^{\mathrm{T}}\right)^{\frac{1}{2}}
\end{aligned}$$

calculate the distance between $p=(2,3,5)$ and $q=(4,7,8)$. As w use $(1,1.5,2.5)$ and as M use both of the following:

$$M_1=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix} \qquad M_2=\begin{pmatrix}1&0.9&0.7\\0.9&1&0.8\\0.7&0.8&1\end{pmatrix}$$

Given 5 pictures as in Figure 1 with 36 pixels each.

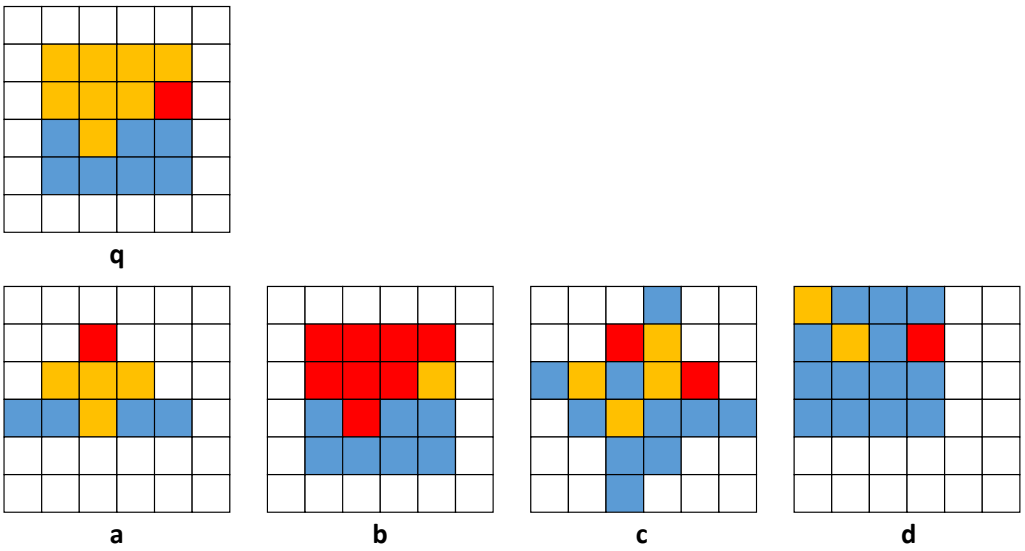

Figure 1: 6×6 pixel pictures

(a) Extract from each picture a color histogram with the bins *red*, *orange*, and *blue* (the white pixels are ignored).

(b) Which pictures are most similar to the query q, using Euclidean distance? Give a ranking according to similarity to q.

(c) The results are not entirely satisfactory. What could you change in the feature extraction or in the distance function to get better results? Report the improved feature extraction and features or the improved distance function.

Exercise 4-3 Visualization of Distance functions

Brainstorm on how you could visualize the behavior of distance measures, and how you could implement a visualization tool. If you got some ideas: well, go an implement them!